

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平9-105748

(43) 公開日 平成9年(1997)4月22日

(51) Int.Cl. ⁶	識別記号	序内整理番号	FI	技術表示箇所
G 0 1 N 33/50			G 0 1 N 33/50	P
G 0 6 F 17/30			G 0 6 F 15/403	3 5 0 A
// C 1 2 N 15/09		9162-4B	C 1 2 N 15/00	A

審査請求 未請求 請求項の数10 O L (全 10 頁)

(21) 出願番号 特願平7-265157

(22) 出願日 平成7年(1995)10月13日

(71) 出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72) 発明者 平岡 進

東京都国分寺市東恋ヶ窪1丁目280番地

株式会社日立製作所中央研究所内

(72) 発明者 永井 啓一

東京都国分寺市東恋ヶ窪1丁目280番地

株式会社日立製作所中央研究所内

(72) 発明者 西川 哲夫

東京都国分寺市東恋ヶ窪1丁目280番地

株式会社日立製作所中央研究所内

(74) 代理人 弁理士 小川 勝男

最終頁に続く

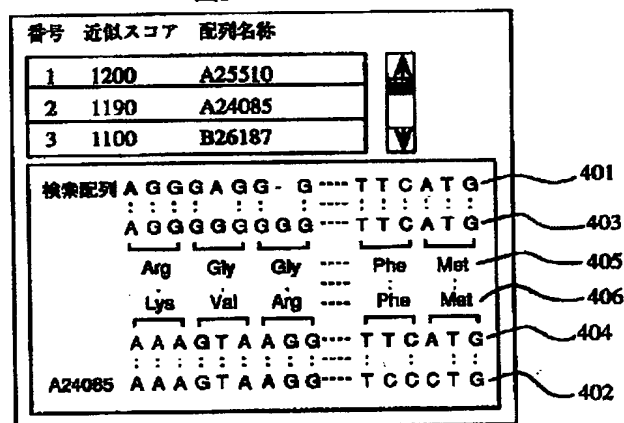
(54) 【発明の名称】 DNA配列比較方法

(57) 【要約】

【課題】 翻訳されるアミノ酸配列が互いに類似するDNA配列を検索することができるDNA配列比較方法を提供する。

【解決手段】 検索DNA配列とデータベース中の既知DNA配列に対して、3文字毎にアミノ酸に翻訳するための中間DNA配列を作成し、中間DNA配列を翻訳してアミノ酸配列を作成する。検索DNA配列を翻訳するための中間DNA配列と検索DNA配列を対応させて表示し、既知DNA配列を翻訳するための中間DNA配列と既知DNA配列を対応させて表示する。中間DNA配列の3文字とアミノ酸を対応させてアミノ酸配列を表示する。対応するアミノ酸配列の間、検索DNA配列と対応する中間DNA配列の間、および既知DNA配列と対応する中間DNA配列の間に類似度を示す記号列を表示する。

図3



【特許請求の範囲】

【請求項1】第一と第二のDNA配列の塩基配列を比較するDNA配列比較方法において、前記第一と第二のDNA配列を所定の長さの塩基群に分割し、分割された塩基群をアミノ酸に翻訳し、前記第一のDNA配列から翻訳されたアミノ酸を第一のアミノ酸配列として、前記第二のDNA配列から翻訳されたアミノ酸を第二のアミノ酸配列として、前記第一のアミノ酸配列と前記第二のアミノ酸配列とを並置して、前記塩基群とともに表示することを特徴とするDNA配列比較方法。

【請求項2】前記所定の長さの塩基群は、挿入、または欠失を含み、前記挿入、または前記欠失を表示することを特徴とする請求項1に記載のDNA配列比較方法。

【請求項3】前記第一と第二のDNA配列を所定の方向に塩基を順次シフトして、前記塩基群に分割し、前記第一と第二のDNA配列から翻訳された前記第一と第二のアミノ酸配列の間で、それぞれのアミノ酸について類似度を積算し、積算結果が最大になるように、前記第一と第二のDNA配列を所定の方向に塩基を順次シフトして、前記第一と第二のアミノ酸配列を選択することを特徴とする請求項1に記載のDNA配列比較方法。

【請求項4】前記第二のDNA配列が、DNA配列データベースから選択された既知DNA配列であることを特徴とする請求項1に記載のDNA配列比較方法。

【請求項5】所定の塩基長の固定長DNA配列とDNA配列データベース内の既知DNA配列とを所定の方向に塩基を順次シフトして、3塩基長からなる塩基群に分割し、前記既知DNA配列と前記固定長DNA配列との分割された前記塩基群のそれぞれをアミノ酸に翻訳し、前記既知DNA配列と前記固定長DNA配列とから翻訳されたアミノ酸からなるアミノ酸配列の間でそれぞれのアミノ酸について類似度を積算し、前記既知DNA配列について、積算結果の最大値を前記固定長DNA配列に対応した構成要素とするスコア表を作成し、前記第一のDNA配列に対して前記既知DNA配列毎に、第一のDNA配列を前記所定の塩基長に分割し、前記既知DNA配列と第一のDNA配列との所定の塩基長毎の類似度の和を前記スコア表の類似度を参照して求め、求めた類似度の和に対応させて順次前記既知DNA配列を表示し、求めた類似度の和の高い前記既知DNA配列を前記第二のDNA配列とすることを特徴とする請求項1に記載のDNA配列比較方法。

【請求項6】前記第一と第二のDNA配列を第一のフォントで表示し、前記第一と第二のアミノ酸配列を第二のフォントで表示することを特徴とする請求項1に記載のDNA配列比較方法。

【請求項7】第一と第二のDNA配列の塩基配列を比較するDNA配列比較方法において、前記第一のDNA配列から、アミノ酸に翻訳するために3塩基ずつに区分された第一の中間DNA配列を作成し、前記第二のDNA

配列から、アミノ酸に翻訳するために3塩基ずつに区分された第二の中間DNA配列を作成し、区分された前記第一と第二の中間DNA配列をそれぞれの3塩基ごとにアミノ酸に翻訳し、前記第一と第二の中間DNA配列のそれぞれの3塩基ごとに対応させて翻訳された前記アミノ酸のそれぞれを並置して、前記第一の中間DNA配列から翻訳されたアミノ酸を第一のアミノ酸配列として、前記第二の中間DNA配列から翻訳されたアミノ酸を第二のアミノ酸配列として、表示し、前記第一のDNA配列と前記第一の中間DNA配列の間の第一の類似度を求め、前記第二のDNA配列と前記第二の中間DNA配列の間の第二の類似度を求め、前記第一と第二のアミノ酸配列の間の第三の類似度を求め、前記第一、第二および第三の類似度から所定の関数を用いて得られるパラメータが最大となるように前記第一と第二の中間DNA配列および前記第一と第二のアミノ酸配列を選択することを特徴とするDNA配列比較方法。

【請求項8】前記第二のDNA配列が、DNA配列データベースから選択された既知DNA配列であることを特徴とする請求項7に記載のDNA配列比較方法。

【請求項9】DNA配列データベース内の既知DNA配列から、アミノ酸に翻訳するために3塩基ずつに区分された第三の中間DNA配列を作成し、所定の塩基長の固定長DNA配列から、アミノ酸に翻訳するために3塩基ずつに区分された第四の中間DNA配列を作成し、前記第三と第四の中間DNA配列の区分されたそれぞれの3塩基ごとにアミノ酸に翻訳し、前記第三と第四の中間DNA配列のそれぞれの3塩基ごとに対応させて翻訳された前記アミノ酸のそれぞれを並置して、前記第三の中間DNA配列から翻訳されたアミノ酸を第三のアミノ酸配列として、前記第四の中間DNA配列から翻訳されたアミノ酸を第四のアミノ酸配列として、前記既知DNA配列と前記第三の中間DNA配列との間の第四の類似度を求め、前記固定長DNA配列と前記第四の中間DNA配列との間の第五の類似度を求め、前記第三と第四のアミノ酸配列の間の第六の類似度を求め、前記既知DNA配列について、前記第四、第五および第六の類似度から所定の関数を用いて得られるパラメータの最大値を前記固定長DNA配列に対応した構成要素とするスコア表を作成し、前記第一のDNA配列に対して前記既知DNA配列毎に、第一のDNA配列を前記所定の塩基長に分割し、前記既知DNA配列と第一のDNA配列との所定の塩基長毎の類似度の和を前記スコア表の類似度を参照して求め、求めた類似度の和に対応させて順次前記既知DNA配列を表示し、求めた類似度の和の高い前記既知DNA配列を前記第二のDNA配列とすることを特徴とする請求項7に記載のDNA配列比較方法。

【請求項10】前記第一と第二のDNA配列を第一のフォントで表示し、前記第一と第二のアミノ酸配列を第二のフォントで表示することを特徴とする請求項7に記載

のDNA配列比較方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、DNA配列比較方法に関し、特にDNA配列の照合に適したDNA配列比較方法に関する。

【0002】

【従来の技術】近年、遺伝子工学の発展によりDNA配列に関するデータが急増している。DNAは生物の設計図とすることができる物質であり、生物の機能は全てDNA中に記述されている。DNAはA、C、G、Tの文字で代表される4種類の物質が並んだ構造となっており、A、C、G、Tだけの文字列であるDNA配列として表すことができる。

【0003】生体中ではDNAの一部が翻訳されて蛋白となり生物機能を担っている。重要な生物機能を担っている蛋白を見つけることが出来れば薬として用いることも可能である。インターフェロン、インスリンなどは薬として製造されるようになった蛋白の例である。蛋白構造全体がわからない場合でも薬物などと結合する結合サイトがわかれば、その蛋白が関係する生物機能を制御する薬の基本構造を決定することができる。蛋白は20種類のアミノ酸が一列に並んだ構造となっており、DNAと同様にその構造はアミノ酸配列として文字列で表すことができる。DNA配列からアミノ酸配列への翻訳はコドンと呼ばれるDNAの3文字が一組となり行われる。図6に示すようにコドンからアミノ酸への翻訳規則は解析されており、コドンからアミノ酸への翻訳は意図的に行われる。DNA配列の翻訳方向と翻訳開始点がわかればDNA配列から3文字ずつコドンを取り出すことで生物中で行われているのと同様にDNAから蛋白の翻訳を行うことができる。ただしDNA配列中に挿入、欠失などの測定誤りがあった場合には、途中からコドンの枠がずれてしまうため蛋白への翻訳を誤ってしまう。

【0004】DNA配列の翻訳方向と翻訳開始点を示す規則は完全にはわかっていない。そのためコドンからアミノ酸への翻訳と異なりDNA配列中の翻訳領域の決定は容易には行われていない。またコドン枠がずれないようにほとんど誤りなくDNA配列を決定するためには配列決定作業が増加してしまう。そのためDNA配列が決定されても対応する蛋白配列の決定が容易に行われないことも多く、未知のDNA配列から未知の蛋白を発見することは容易ではない。未知の蛋白の発見は新薬の開発に重要である。

【0005】DNA配列から翻訳領域を決定するなどして蛋白を決定する手法には、翻訳領域推定ソフト、蛋白測定、cDNA測定、ハイブリダイゼーション、類似DNA配列検索などがある。翻訳領域推定ソフトはこれまでわかっている知識を総合してDNA配列から蛋白翻訳領域を推定するソフトである。蛋白翻訳に関する知識が

完全ではないため精度はあまり高くない。蛋白測定は生物中で合成されている蛋白を直接測定する方法であり、感度が高くなくアミノ酸配列測定も容易でないため、大量の生物試料が必要とされる。

【0006】cDNA測定はDNAから蛋白へ翻訳される過程の中間物質をDNAに変換して測定する方法である。蛋白と異なりDNAは増幅可能なため、cDNA測定は蛋白測定に比べて感度は高い。それでも微量な蛋白に対応するcDNAは大量な蛋白に対応するcDNAの中に混合し埋もれてしまうため、特殊な蛋白を測定するためには特殊な生物試料からcDNAを抽出するか、ハイブリダイゼーション、DNA配列相同性検索などを用いて特定のcDNAを選択する必要がある。

【0007】ハイブリダイゼーションはすでに手元にあるDNAに対して類似しているDNAをDNA混合試料または長いDNAから選択する技術である。手元にあるDNAに相補的なDNAを合成し、これに結合するDNAを選択することで類似DNAを選択することができる。

20 【0008】DNA配列相同性検索（ホモロジーサーチ）はハイブリダイゼーションが生物学的手法で行っていることを情報処理的手法で行うものである。あらかじめ全てのDNA配列を測定しておく必要があるが、ハイブリダイゼーションと異なり相補的結合を行うようなDNAばかりでなく、異なった基準で類似しているDNA配列を選択することができる。

30 【0009】相同性検索を行う最も基本的な方法として、問い合わせ配列とデータベース内の各配列との間でスミスとウォーターマンにより提案されたダイナミックプログラミング法によるアライメントを行い、高いスコア順に各配列を表示する方法がある。ダイナミックプログラミング法によるアライメントに関してはジャーナル・オブ・モレキュラー・バイオロジー、147（1981年）195頁から197頁（Journal of Molecular Biology、147（1981）pp195-197）に記載されている。

40 【0010】ダイナミックプログラミング法による配列1（CAGTGACT）と配列2（CACTGCTG）のアライメントを図7を用いて説明する。ダイナミックプログラミング法によるアライメントでは2次元メッシュのX、Y方向に添ってそれぞれ2本の配列を置き、メッシュの各点をノードとして、ノード間には縦、横、斜めの3方向の経路を考えた時に任意の2つのノード間を左上から右下に向かう最適経路を求める。縦、横のアー

5

うしのアライメントにおいて一般的に用いられているスコアは、挿入・欠失のスコアは n 文字の挿入・欠失に対して $-4n-8$ 点、一致した1文字のスコアは4点、異なっている1文字のスコアは-3点である。例えば図7に示した経路でのスコアは9点である。

【0011】相同性検索プログラムであるFASTAについては、アカデミックプレス (Academic Press) より発行されたドゥーリトル (Doolittle) 編集によるメソツツ・イン・エンザイモロジー、183 (1990年) 第63頁から98頁 (Methods in Enzymology, 183 (1990) pp 63-98) に記載されている。FASTAではダイナミックプログラミング法によるスコアよりも少ない計算量で求められる $initn$ 、 $initl$ と呼ばれるスコアを求めている。これらのスコアは完全に一致する一定長の部分文字列を探し出し、それらを繋ぎあわせて求める。FASTAではこれらのスコアが高い順に配列を表示しており、さらに上位の配列に対してはダイナミックプログラミング法によるスコアを計算しOPTという名で表示している。

【0012】また、相同性検索プログラムであるBLASTについては、ジャーナル・オブ・モレキュラー・バイオロジー、215 (1990年) 第403頁から410頁 (Journal of Molecular Biology, 215 (1990) pp 403-410) に記載されている。BLASTでもダイナミックプログラミング法によるスコアよりも少ない計算量で求められるスコアを求めている。BLASTではFASTAにおける $initl$ よりも単純化したスコアを用いている。BLASTではBLAST独自のスコア計算のみであり、ダイナミックプログラミング法によるスコアの計算は行っていない。

【0013】スミスとウォーターマンによるダイナミックプログラミング法、FASTA、BLASTはいずれもDNA配列をDNA配列データベースに対して検索する方法であり、DNA配列どうしを比較している。これらのプログラムはアミノ酸配列どうしを比較することで、アミノ酸配列をアミノ酸配列データベースに対して検索することも可能である。アミノ酸配列の比較で用いられているスコアを図8に示す。

【0014】これに対してDNA配列をアミノ酸配列データベースに対して検索するBLASTX、そしてアミノ酸配列をDNA配列データベースに対して検索するTFASTA、TBLASTNが存在する。これらは比較する際にDNA配列を方向と翻訳開始点を変えた6通りのアミノ酸配列に翻訳し、アミノ酸配列どうしで比較している。

【0015】

【発明が解決しようとする課題】従来の配列比較、検索プログラムでは、DNA配列とアミノ酸配列の比較にお

6

いてDNA配列の翻訳を行っていたが、DNA配列どうしの比較においては両DNA配列をアミノ酸に翻訳し比較することを行っていなかった。DNA配列のアミノ酸翻訳領域はたとえ翻訳されたアミノ酸がわからない場合でも、DNA配列そのままと比較するよりもアミノ酸配列に翻訳して比較することが望ましい。類似機能を持つ蛋白は類似アミノ酸配列部分を持ち、その部分が共通する重要な働きをしているからである。

【0016】コドンからアミノ酸への翻訳は多対一であるため異なるDNA配列でも同じ蛋白に翻訳されることがある。さらにコドンが完全に異なり異なるアミノ酸に翻訳される場合でも、それらのアミノ酸の性質が類似していることがある。言い替えれば翻訳されたアミノ酸配列が類似していても元のDNA配列が類似しているとは限らない。アミノ酸配列に翻訳して比較することで翻訳されたアミノ酸配列が類似しているDNA配列と類似している部分を選択することができる。類似アミノ酸配列に翻訳される部分はアミノ酸配列翻訳領域であり、蛋白の共通機能部分であることが推定できる。

【0017】従来のDNA配列検索ではDNA配列そのものが類似している順番にDNA配列データベースを並べていた。そのため翻訳されたアミノ酸が類似しているDNA配列は多くの偶然似ているDNA配列に埋もれてしまっていて見つけられない点に問題がある。DNA配列どうしを一对一で比較する場合にも多くの偶然類似しているDNA部分配列の中に埋もれてしまい、翻訳されたアミノ酸配列が類似しているDNA配列部分を見つけることは困難である点に問題がある。

【0018】本発明の目的は、DNA配列から翻訳されたアミノ酸配列を比較することで、翻訳されるアミノ酸配列が互いに類似するDNA配列をもれなく検索することができるDNA配列比較方法を提供することにある。

【0019】

【課題を解決するための手段】上記目的を達成するために本発明では、比較するDNA配列をアミノ酸配列に翻訳し、翻訳されたアミノ酸配列を比較することを特徴とする。翻訳したアミノ酸配列の類似部分配列を対応させてアミノ酸配列を並置表示し、DNA配列の各コドンに翻訳された各アミノ酸に対応させて表示する。DNA配列からアミノ酸配列への翻訳と、アミノ酸配列どうしの比較結果を表示する。DNA配列をアミノ酸配列に翻訳し、翻訳されたアミノ酸を比較することで、翻訳されたアミノ酸配列が類似しているDNA配列を見つけることができ、翻訳されたアミノ酸配列が互いに類似するDNA配列をもれなく検索することができる。

【0020】以下、より詳細に本発明の特徴を説明すると、第一と第二のDNA配列を所定の長さの塩基群に分割し、分割された塩基群をアミノ酸に翻訳し、第一のDNA配列から翻訳されたアミノ酸を第一のアミノ酸配列として、第二のDNA配列から翻訳されたアミノ酸を第

二のアミノ酸配列として、第一のアミノ酸配列と第二のアミノ酸配列とを並置して、塩基群とともに表示することを特徴とする。所定の長さの塩基群は、挿入、または欠失を含み、挿入、または欠失を表示する。第一と第二のDNA配列を所定の方向に塩基を順次シフトして、塩基群に分割し、第一と第二のDNA配列から翻訳された第一と第二のアミノ酸配列の間で、それぞれのアミノ酸について類似度を積算し、積算結果が最大になるように、第一と第二のDNA配列を所定の方向に塩基を順次シフトして、第一と第二のアミノ酸配列を選択する。第二のDNA配列が、DNA配列データベースから選択された既知DNA配列であってもよい。

【0021】所定の塩基長の固定長DNA配列とDNA配列データベース内の既知DNA配列とを所定の方向に塩基を順次シフトして、3塩基長からなる塩基群に分割し、既知DNA配列と固定長DNA配列との分割された塩基群のそれぞれをアミノ酸に翻訳し、既知DNA配列と固定長DNA配列とから翻訳されたアミノ酸からなるアミノ酸配列の間でそれぞれのアミノ酸について類似度を積算し、既知DNA配列について、積算結果の最大値を固定長DNA配列に対応した構成要素とするスコア表を作成し、第一のDNA配列に対して既知DNA配列毎に、第一のDNA配列を所定の塩基長に分割し、既知DNA配列と第一のDNA配列との所定の塩基長毎の類似度の和をスコア表の類似度を参照して求め、求めた類似度の和に対応させて順次既知DNA配列を表示し、求めた類似度の和の高い既知DNA配列を第二のDNA配列としてもよい。第一と第二のDNA配列を第一のフォントで表示し、第一と第二のアミノ酸配列を第二のフォントで表示する。

【0022】また、本発明では、第一のDNA配列から、アミノ酸に翻訳するために3塩基ずつに区分された第一の中間DNA配列を作成し、第二のDNA配列から、アミノ酸に翻訳するために3塩基ずつに区分された第二の中間DNA配列を作成し、区分された第一と第二の中間DNA配列をそれぞれの3塩基ごとにアミノ酸に翻訳し、第一と第二の中間DNA配列のそれぞれの3塩基ごとに対応させて翻訳されたアミノ酸のそれぞれを並置して、第一の中間DNA配列から翻訳されたアミノ酸を第一のアミノ酸配列として、第二の中間DNA配列から翻訳されたアミノ酸を第二のアミノ酸配列として、表示し、第一のDNA配列と第一の中間DNA配列の間の第一の類似度を求め、第二のDNA配列と第二の中間DNA配列の間の第二の類似度を求め、第一と第二のアミノ酸配列の間の第三の類似度を求め、第一、第二および第三の類似度から所定の関数を用いて得られるパラメータが最大となるように第一と第二の中間DNA配列および第一と第二のアミノ酸配列を選択することを特徴とする。第二のDNA配列が、DNA配列データベースから選択された既知DNA配列であってもよい。

【0023】DNA配列データベース内の既知DNA配列から、アミノ酸に翻訳するために3塩基ずつに区分された第三の中間DNA配列を作成し、所定の塩基長の固定長DNA配列から、アミノ酸に翻訳するために3塩基ずつに区分された第四の中間DNA配列を作成し、第三と第四の中間DNA配列の区分されたそれぞれの3塩基ごとにアミノ酸に翻訳し、第三と第四の中間DNA配列のそれぞれの3塩基ごとに対応させて翻訳されたアミノ酸のそれぞれを並置して、第三の中間DNA配列から翻訳されたアミノ酸を第三のアミノ酸配列として、第四の中間DNA配列から翻訳されたアミノ酸を第四のアミノ酸配列として、既知DNA配列と第三の中間DNA配列との間の第四の類似度を求め、固定長DNA配列と第四の中間DNA配列との間の第五の類似度を求め、第三と第四のアミノ酸配列の間の第六の類似度を求め、既知DNA配列について、第四、第五および第六の類似度から所定の関数を用いて得られるパラメータの最大値を固定長DNA配列に対応した構成要素とするスコア表を作成し、第一のDNA配列に対して既知DNA配列毎に、第一のDNA配列を所定の塩基長に分割し、既知DNA配列と第一のDNA配列との所定の塩基長毎の類似度の和をスコア表の類似度を参照して求め、求めた類似度の和に対応させて順次既知DNA配列を表示し、求めた類似度の和の高い既知DNA配列を第二のDNA配列としてもよい。第一と第二のDNA配列を第一のフォントで表示し、第一と第二のアミノ酸配列を第二のフォントで表示する。

【0024】

【発明の実施の形態】本発明を検索DNA配列に対するDNA配列データベース検索に応用した実施例について詳述する。あらかじめDNA配列データベースに対してスコア表を作成しておく。スコア表はDNA配列データベース内の各DNA配列について、所定長の固定長DNA配列に対応したアミノ酸配列に翻訳した時のスコア（類似度）を、構成要素とする表である。図1を用いてDNA配列データベース内のDNA配列と固定長DNA配列とをアミノ酸配列に翻訳した時のスコアの計算方法を説明する。

【0025】DNA配列データベース内のDNA配列をDNA配列301、固定長DNA配列をDNA配列302とする。DNA配列301から中間DNA配列303を生成し、DNA配列302から中間DNA配列304を生成する。中間DNA配列303、304はアミノ酸配列に翻訳するために3塩基ずつに区分されている。3塩基ずつ翻訳していくことで、中間DNA配列303からアミノ酸配列305が生成され、中間DNA配列304からアミノ酸配列306が生成される。

【0026】DNA配列301と中間DNA配列303をダイナミックプログラミング法を用いてアライメント（並置）し、スコア307を求め、DNA配列302と

中間DNA配列304をダイナミックプログラミング法を用いてアライメントし、スコア308を求め、アミノ酸配列305とアミノ酸配列306をダイナミックプログラミング法を用いてアライメントし、スコア309を求める。スコア307、308、309は用いる中間DNA配列303、304によって様々な値をとる。適当な中間DNA配列303、304を用いて、スコア307、308、309から所定の関数を用いて得られるパラメータの最大値を求める。この最大値がDNA配列301とDNA配列302をアミノ酸配列に翻訳したときの類似性を示すスコアとなる。この値をスコア表に格納する。

【0027】DNA配列データベース検索の際は、検索DNA配列を固定長DNA配列と同じ塩基長に分割し、DNA配列データベース内の各DNA配列と検索DNA配列との固定長DNA配列の塩基長毎のスコアの和をスコア表のスコアを参照して求め近似スコアとする。前記近似スコアに対応させて順次DNA配列データベース内の各DNA配列の配列名称を近似スコアの値の大きい順に近似スコアと共に表示する。

【0028】表示されたDNA配列データベース内の各DNA配列の中からDNA配列を選択し、選択されたDNA配列と検索DNA配列とをアミノ酸に翻訳したときの類似性を示すスコアを求める計算を行う。このスコアの求め方を図2を用いて以下で説明する。

【0029】DNA配列データベース内のDNA配列をDNA配列402、検索DNA配列をDNA配列401とする。DNA配列401から中間DNA配列403を生成し、DNA配列402から中間DNA配列404を生成する。中間DNA配列403、404はアミノ酸配列に翻訳するために3塩基ずつに区分されている。3塩基ずつ翻訳していくことで、中間DNA配列403からアミノ酸配列405が生成され、中間DNA配列404からアミノ酸配列406が生成される。DNA配列401と中間DNA配列403をダイナミックプログラミング法を用いてアライメント（並置）し、スコア407を求め、DNA配列402と中間DNA配列404をダイナミックプログラミング法を用いてアライメントし、スコア408を求め、アミノ酸配列405とアミノ酸配列406をダイナミックプログラミング法を用いてアライメントし、スコア409を求める。スコア407、408、409は用いる中間DNA配列403、404によって様々な値をとる。中間DNA配列403、404には、スコア407、408、409から所定の関数を用いて得られるパラメータが最大値となるDNA配列を選択する。パラメータの最大値はDNA配列401とDNA配列402をアミノ酸配列に翻訳したときの類似性を示すスコアである。

【0030】本実施例を用いた検索結果を図3に示す。近似スコアに対応させて順次DNA配列データベース内

の各DNA配列の配列名称を近似スコアの値の大きい順に近似スコアと共に表示する。DNA配列401（検索DNA配列）と中間DNA配列403を対応させて表示し、DNA配列402（DNA配列データベース内のDNA配列（A24085））と中間DNA配列404とを対応させて表示し、アミノ酸配列405とアミノ酸配列406を対応させて表示し、対応するDNAの塩基が一致していれば記号“:”、類似していれば記号“.”、一致も類似もしていなければ記号としてブランクを対応するDNAの間に表示し、対応するアミノ酸が一致していれば記号“:”、類似していれば記号“.”、一致も類似もしていなければ記号としてブランクを対応するアミノ酸の間に表示する。

【0031】スコア表に用いる固定長配列の長さは長い方が望ましい。固定長DNA配列の長さが短い場合には、スコア表のほとんど全ての構成要素がおなじ値になり、スコア表から計算される近似スコアもほとんど同じ値になり、結果として検索DNA配列と翻訳したアミノ酸が似ているデータベース内の配列を選ぶことが出来ないからである。例えば固定長DNA配列の長さとして6文字を用いた場合には、翻訳されるアミノ酸配列の長さは2文字である。データベース中のDNA配列を翻訳すると、かなりの頻度で固定長DNA配列から翻訳されたアミノ酸とデータベース中のDNA配列から翻訳されたアミノ酸が同じになることが起こり得る。これによりスコア表のほとんどの構成要素の値は完全に一致した場合のスコアとなる。そこでスコア表に用いる固定長配列の長さとしては7文字以上が望ましい。

【0032】スコアの関数としては最も単純には3つのスコアの和が考えられる。単純な和以外にも自乗なども可能である。そしてDNA配列どうしのスコアとアミノ酸配列どうしのスコアも、従来用いられているスコアそのものに限定する必要はない。例えばDNAどうしのスコアを半分にすれば、より大きいDNA配列測定誤差を許容した比較が可能となる。DNA配列決定方法によってDNA配列測定誤差は異なるため、それに応じてスコアの関数、DNA配列どうしのスコア、そしてアミノ酸配列どうしのスコアを使い分けることが望ましい。

【0033】中間DNA配列403と中間DNA配列404は計算時間が問題とならない場合には、あらゆるDNA配列を生成することが望ましい。この場合、DNA配列どうしの直接比較における場合と同様にダイナミックプログラミング法によるスコア計算が可能である。

【0034】図4を参照して、中間DNA配列を用いてアミノ酸に翻訳し比較するときのダイナミックプログラミング法によるスコア計算方法を以下に示す。2次元メッシュのY方向に沿ってDNA配列501を置き、X方向に沿ってDNA配列502を置いた。図4ではメッシュの各点がノード101～113を示し、格子間を結ぶ線分がアーク201～206を示す。各ノードを左上か

ら右下につないでいった経路がアライメント（並置）に対応しており、アークは経路を分割したときの最小単位である。そして各ノードにおけるスコアはDNA配列501とDNA配列502のそこまでの部分配列どうしを比較したときのスコアに対応している。

【0035】ダイナミックプログラミング法では、メッシュ中のあらゆる経路のスコアを求め最大値を決定する計算を、左上から右下に向かって各ノードのスコアを順番に求めていく計算に置き換えて計算の高速化を果たしている。各ノードのスコアはそのノードで終る経路のスコアの最大値を示している。各ノードのスコアは、そのノードと単独のアークで結ばれる上、左または左上の各ノードのスコアとそこからのアークのスコアを加えたスコアを求め、それらの最大値を求めることで行われる。ダイナミックプログラミング法では全てのノードのスコアを求めた後、全てのノードのスコアの中の最大値を求めることでDNA配列501とDNA配列502の間のスコアを決定することができる。

【0036】DNA配列を直接比較する従来のダイナミックプログラミング法では、ノード101で終るアークは、DNA挿入に対応するアーク201、DNA欠失に対応するアーク202、DNAどうしの比較に対応するアーク203である。ノード102のスコアとアーク201のスコアの和、ノード105のスコアとアーク202のスコアの和、ノード109のスコアとアーク203のスコアの和、および0の中の最大値がノード101におけるスコアとなる。

【0037】DNA配列を中間DNA配列を用いてアミノ酸配列に翻訳して比較するときのダイナミックプログラミング法では、従来のDNA配列を直接比較するときのダイナミックプログラミング法とは異なったアークを用いる必要がある。ノード101に達するアークには、コドンの境界へのDNA挿入に対応するアーク201、コドンの境界へのDNA欠失に対応するアーク202、翻訳したアミノ酸の挿入に対応するアーク204、翻訳したアミノ酸の欠失に対応するアーク205、翻訳したアミノ酸どうしの比較に対応するアーク206がある。DNA配列の直接比較におけるダイナミックプログラミング法と異なりDNAどうしの比較に対応するアーク203は用いない。そしてアミノ酸への翻訳では中間DNA配列中のコドン considering することで、コドン中のDNAの置換、挿入、欠失を含めて行う。

【0038】例えば、アーク206はDNA配列501中のDNA3文字AACとDNA配列502中のDNA4文字ATCTを比較している。DNA配列501に対応する中間DNA配列中のコドンをAAC、DNA配列502に対応する中間DNA配列中のコドンをTCTとしたときのアーク206のスコアは、DNA配列501中のDNA3文字AACとDNA配列501に対応する中間DNA配列中のコドンAACとの間のスコアと、D

NA配列502中のDNA4文字ATCTとDNA配列502に対応する中間DNA配列中のコドンTCTとの間のスコアと、DNA配列501に対応する中間DNA配列中のコドンAACを翻訳したアミノ酸AsnとDNA配列502に対応する中間DNA配列中のコドンTCTを翻訳したアミノ酸Serとの間のスコアの所定の関数の値となる。そしてアーク206のスコアは、DNA配列501に対応する中間DNA配列とDNA配列502に対応する中間DNA配列に対して全ての塩基の組合せでできるコドンに対して求めた前記の所定の関数の最大値である。

【0039】中間DNA配列中のコドンは3文字だが、中間DNA配列中のコドンと比較するDNA配列中のDNAは3文字である必要はない。アーク206はその一例である。図4に示した翻訳したアミノ酸の比較に対応するアークは一例であり、ノード101の上に位置するノード102、103、104などからノード101へのアークすべてが翻訳したアミノ酸の挿入に対応するアークであり、ノード101の左に位置するノード105、106、107、108などからノード101へのアークすべてが翻訳したアミノ酸の欠失に対応するアークであり、ノード101の左上に位置するノード109、110、111などからノード101へのアークすべてが翻訳したアミノ酸どうしの比較に対応するアークとなる。これらすべてのアークに対して、アークの始まりのノードのスコアとアークのスコアの和を求め、最大値を求めることでノード101のスコアが求められる。

【0040】図4に示した中間DNA配列を用いてアミノ酸翻訳して比較するダイナミックプログラミング法による計算は、全ての中間DNA配列を考慮することによってDNA配列測定における測定誤りを考慮しており、DNA配列の途中にDNA挿入が入ってコドンの読み枠がずれた場合にも対応できる。

【0041】ダイナミックプログラミング法では計算時間が長く問題となる場合には、生成する中間DNA配列を制限することも可能である。翻訳方向と読み枠を固定した中間DNA配列403として、DNA配列401を順方向と逆方向にそれぞれ3種類の読み枠で固定したDNA配列を考える。中間DNA配列404も同様に6種類考える。この場合中間DNA配列403と中間DNA配列404の組み合わせは36通りであり、全ての組み合わせでスコア計算を行っても計算時間は長くない。

【0042】図5は、本発明の一装置構成を示す図である。図5を用いて、計算処理手順を示す。スコア表記憶器1には上記方法で計算したデータベース内の各DNA配列と一定長のあらゆるDNA配列とのアミノ酸配列に翻訳した時の類似性を示すスコアを、スコア表として作成しておく。検索の際は始めにデータベース各配列の近似スコアを記憶しておく近似スコア記憶器2と一時スコア記憶器8をすべて0にリセットし、検索配列を検索配

列記憶器3に記憶する。

【0043】次に、カウンタ4を0から増分させていく。カウンタ4の出力は上位桁は部分配列取り出し器5に入力され部分配列切り出し部分を指定し、下位桁はスコア表記憶器1と近似スコア記憶器2と一時スコア記憶器8に入力されデータベース配列番号を指定する。指定された各配列の近似スコアと一時スコアはそれぞれスコア記憶器14と15に記憶される。検索配列記憶器3に接続された部分配列取り出し器5によって検索配列は一定長k文字づつ切り出され、カウンタ4によって指定された部分配列が部分配列記憶器6に記憶される。部分配列記憶器6の出力はスコア表記憶器1に入力されカウンタ4で指定されたデータベース内の各配列とのダイナミックプログラミング法によるスコアがスコア表記憶器1から出力される。スコア表記憶器1の出力は加算器7を用いてスコア記憶器14のスコアと加算される。加算器7の出力は比較器9で0と比較され大きい値が出力され、スコア記憶器14に書き戻される。比較器9の出力は、スコア記憶器15のスコアと比較器10で比較され大きい値が再びスコア記憶器15に書き戻される。その後スコア記憶器14と15の内容はそれぞれカウンタ4で指定される一時スコア記憶器8と近似スコア記憶器2の対応する部分に書き戻される。

【0044】カウンタ4により検索配列から切り出されたすべての部分配列に対してすべてのデータベース配列が走査しつくされたら、近似スコア記憶器2の各スコアをソーター11によって大きいスコアの順に並べ替え、配列名称記憶器13によって対応する配列名称を求め表示器12に表示する。次に表示されたDNA配列データベース中の配列の中から指定されたDNA配列と検索DNA配列を、アミノ酸配列に翻訳して比較する。

【0045】図3の翻訳比較では、DNA配列と中間DNA配列との並置、中間DNA配列のアミノ酸配列への翻訳、翻訳されたアミノ酸配列どうしの並置を全て表示している。従来のアミノ酸配列どうしの並置と同様にアミノ酸配列の類似部分を表示することによって、共通する蛋白構造の位置、配列、類似度を同時に示している。これは蛋白への薬物結合サイトの発見から薬物設計、そして蛋白の共通機能の推定などに役立てることができる。同時に表示される中間DNA配列のアミノ酸配列への翻訳からは用いられているコドン調べることができる。生物種によって用いられるコドンにはかたよりがあるため、生物種がわかっている場合ここで表示されている翻訳が正しいかどうかを推定することができる。またDNA配列と中間DNA配列との並置からはDNA配列測定誤差の傾向を知ることができる。これはDNA配列測定実験の精度を上げるために役立てられ、測定誤差の大きい部分のDNA配列を再測定するための指針となる。

【0046】本実施例ではDNA配列と中間DNA配列との並置、中間DNA配列のアミノ酸配列への翻訳、

訳されたアミノ酸配列どうしの並置を同時に表示することで、DNA配列測定誤差、アミノ酸翻訳誤り、アミノ酸配列共通構造を一目で総合的に判断できる。本実施例と異なり中間DNA配列、DNA配列と中間DNA配列との並置を省略することも可能である。省略した場合より多くの結果を表示することが可能となる。さらに全くDNA配列の表示がなく翻訳されたアミノ酸配列の並置だけを表示した場合にもDNA配列だけの情報から類似アミノ酸配列を知ることができるため有効である。また複数種類のフォントまたは色が表示可能な表示器の場合には、DNA配列とアミノ酸配列のフォントまたは色に異なるものを用いることが可能である。これによりアミノ酸配列をDNA配列よりも強調し、DNA配列の比較表示の中でもっとも重要であるアミノ酸配列の並置が一目でわかるようにすることができる。

【0047】本実施例で用いている近似スコアは検索DNA配列を分割した部分DNA配列とDNA配列データベース中のDNA配列とのアミノ酸翻訳をして比較したスコアを加算した結果となっている。本実施例ではあらかじめデータベース内の各配列に対して一定長のあらゆる文字列とのアミノ酸翻訳をして比較したスコア表を作成してある。そのため近似スコアはスコア表を参照し和をとるだけで求めることが出来、計算量は小さい。

【0048】本実施例では2つのDNA配列間でアミノ酸に翻訳して比較しているが、本発明は3つ以上のDNA配列間にも適用可能である。多数のDNA配列を本方法で比較することによって翻訳されたアミノ酸配列に共通する構造を見つけ出すことができる。

【0049】

【発明の効果】本発明により、DNA配列に対して翻訳されたアミノ酸配列構造がわからない場合にもアミノ酸配列の共通構造を発見できる。アミノ酸配列の共通構造の発見は、薬物の結合サイトの発見から薬物設計へ応用したり、類似蛋白機能の発見から蛋白設計などに役立つ。

【0050】また、DNA配列のアミノ酸配列への翻訳とアミノ酸配列どうしの並置表示を同時に行うことでDNA配列測定誤差、アミノ酸翻訳誤り、アミノ酸配列共通構造を一目で総合的に判断できる。

【0051】また、DNA配列から翻訳されるアミノ酸配列を比較することで、翻訳されるアミノ酸が互いに類似するDNA配列をもれなく検索することができる。

【図面の簡単な説明】

【図1】DNA配列データベース内のDNA配列と固定長DNA配列とをアミノ酸配列に翻訳してスコアを計算する方法を説明する図。

【図2】選択されたDNA配列と検索DNA配列とをアミノ酸配列に翻訳してスコアを求める方法を説明する図。

【図3】実施例の表示画面を示す図。

15

【図4】中間DNA配列をアミノ酸に翻訳して比較するときの、ダイナミックプログラミング法によるスコア計算方法を説明する図。

【図5】本発明の一装置構成を示す図。

【図6】アミノ酸翻訳コドン表を示す図。

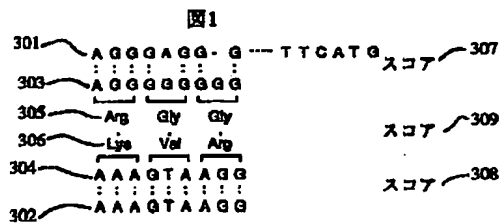
【図7】ダイナミックプログラミング法を説明する図。

【図8】アミノ酸スコア表を示す図。

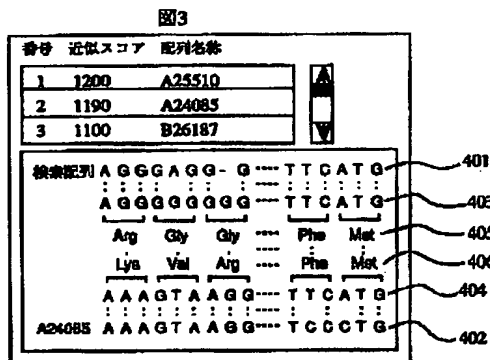
【符号の説明】

1…スコア表記憶器、2…近似スコア記憶器、3…検索

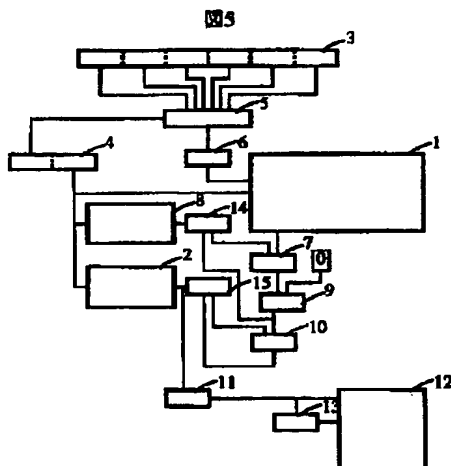
【図1】



【図3】



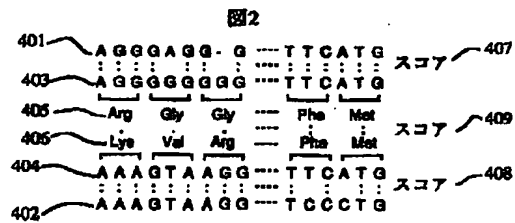
【図5】



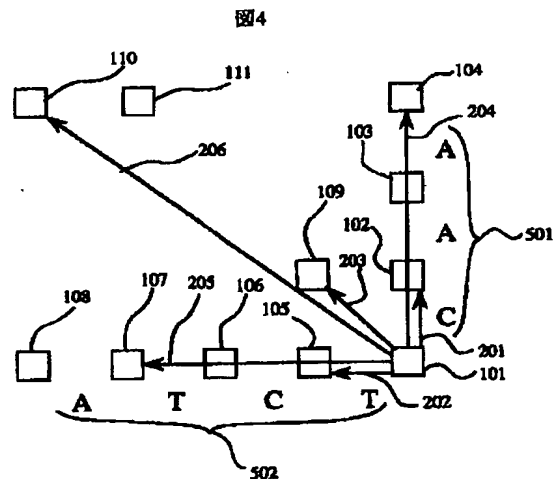
16

配列記憶器、4…カウンタ、5…部分配列取り出し器、6…部分配列記憶器、7…加算器、8…一時スコア記憶器、9、10…比較器、11…ソーター、12…表示器、13…配列名称記憶器、14、15…スコア記憶器、101~113…ノード、201~206…アーク、301、302、401、402、501、502…DNA配列、303、304、403、404…中間DNA配列、305、306、405、406…アミノ酸配列、307~309、407~409…スコア。

【図2】



【図4】



【図6】

図6

1番目 (5'末端)	2番目				3番目 (3'末端)
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	終止	終止	A
	Leu	Ser	終止	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

【図7】

図7

配列1

		C	A	G	T	G	A	C	T
配列2	C	43							
	A		42						
	C			41					
	T				40				
	G					39			
	C						38		
	T							37	
	G								36
	C								
	T								

○：各スコア

【図8】

図8

システイン	C	IS	イオウ置換性																						
セリン	S	0	2	小型																					
トレオニン	T	-2	1	3																					
プロリン	P	-3	1	0	0																				
アラニン	A	-2	1	1	1	2																			
グリシン	G	-3	1	0	-1	1	3																		
アスパラギン	N	-4	1	0	-1	0	0	2	酸性とそのアミド																
アスパラギン酸	D	-5	0	0	-1	0	1	2	4																
グルタミン酸	E	-6	0	0	-1	0	0	1	3	4															
グルタミン	Q	-5	-1	-1	0	0	-1	1	2	2	4														
ヒスチジン	H	-3	-1	-1	0	-1	-2	2	1	1	3	0	塩基性												
アルギニン	R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	5												
リシン	K	-3	0	0	-2	-1	-2	1	0	0	1	0	3	5											
メチオニン	M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-3	0	0	0	疎水性									
イソロイシン	I	-3	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-3	-3	2	5									
ロイシン	L	-6	-3	-2	-3	-2	-6	-3	-4	-3	-7	-2	-3	-3	4	2	5								
バリン	V	-2	-1	0	-1	0	-3	-2	-2	-4	-2	-1	-2	-2	2	4	2	4							
フェニルアラニン	F	-4	-3	-3	-1	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	8	芳香族					
チロシン	Y	0	-3	-3	-3	-3	-6	-2	-4	-4	-6	0	-4	-4	-2	-1	-1	-2	7	10					
トリプトファン	W	-9	-3	-5	-5	-5	-7	-4	-7	-7	-9	-3	2	-1	-4	-5	-2	-6	0	0	17				
		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W				

フロントページの続き

(72)発明者 笠原 直子

東京都国分寺市東恋ヶ窪1丁目280番地

株式会社日立製作所中央研究所内